DOCUMENT RESUME

ED 040 061 | 24 | SE 008 725

AUTHOR	Ferguson, Richard L.
TITLE	Computer-Assisted Criterion-Referenced Testing, Working Paper 49.
INSTITUTION	Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.; Office of Naval Research, Washington, D.C. Psychological Sciences Div.
BUREAU NO	BR-5-0253
PUB DATE	69
NOTE	18p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.00
DESCRIPTORS	*Computer Assisted Instruction, Computers, *Elementary School Mathematics, *Evaluation, Measurement, *Testing
IDENTIFIERS	*Individually Prescribed Instruction

ABSTRACT

Reported on is a study to develop a model for computer-assisted branched testing, measurement in which items are selected on the basis of previous responses and are thus tailored to the competencies of the examinee. The model was developed, implemented and evaluated in the context of an experimental school in Individually Prescribed Instruction (IPI). The sample consisted of seventy-five students in grades one through six at the Oakleaf Elementary School. On two separate occasions each student was given the branched computer test using a teletypewriter. After completing the branched test, students were required to take a paper and pencil test on all remaining objectives not included on the branched test. Thus, a measure of proficiency of every objective in the unit was recorded. Individual test profiles resulting from the study were used to ascertain the validity of the unit's structure. The results showed that criterion-referenced branched testing can be used effectively for all students in an IPI setting. Further, the measures yielded by such a procedure can be at least as valid and reliable as those for a conventional test. (RP)

BR-5-0253
PA 24

UNIVERSITY OF PITTSBURGH - LEARNING R & D CENTER

WORKING PAPER 49

COMPUTER-ASSISTED CRITERION-REFERENCED TESTING

RICHARD L. FERGUSON

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED0 40061

SE 008 795

EDO 40061

COMPUTER-ASSISTED CRITERION-REFERENCED MEASUREMENT[1]

Richard L. Ferguson

Learning Research and Development Center

University of Pittsburgh

Fall, 1969

[1]Support of this study was provided by the United States Office of Education, Department of Health, Education, and Welfare, under provisions of the Cooperative Research Program. Additional support was provided by the Personnel and Training Branch, Psychological Sciences Division, Office of Naval Research.

COMPUTER-ASSISTED CRITERION-REFERENCED MEASUREMENT[1]

Richard L. Ferguson[2]
Learning Research and Development Center
University of Pittsburgh

Accommodating instruction to the specific needs of individuals is paramount among the goals of recent innovations in education. Changes in testing procedures should be a natural outgrowth of attempts to individualize instruction. Accordingly, computer-assisted branched testing reflects one possible direction which such new developments in testing might take.

The purpose of this study was to develop a model for computer-assisted branched testing, measurement in which items are selected on the basis of previous responses and are thus tailored to the competencies of the examinee. The model was developed, implemented and evaluated in the context of an experimental school in Individually Prescribed Instruction (IPI). IPI is a joint project of the University of Pittsburgh's Learning Research and Development Center and the Baldwin-Whitehall School District. The major feature of the project is that prescriptions for instruction are adapted to the individual differences among children.

Initial studies concerned with branched testing have resulted in a cautious optimism regarding its potential for measurement purposes. Numerous studies (Bayroff and Seeley, 1967; Waters, 1964; Hanson and Schwarz, 1968) have reported some initial success with branched tests while others (Cleary et al., 1968; Angloff and Huddleston, 1958) have posed questions as to the merit of

---

[1]Support for this study was provided by the United States Office of Education, Department of Health, Education, and Welfare, under provisions of the Cooperative Research Program. Additional support was provided by the Personnel and Training Branch, Psychological Sciences Division, Office of Naval Research.

[2]The author is indebted to Dr. William Cooley, Dr. Robert Glaser, and Dr. Anthony Nitko for their helpful suggestions throughout the study. Appreciation is further expressed to Dr. Richard Cox and Dr. Grace Lazovik for their assistance.

using such devices for measurement purposes since in many cases short conventional tests could achieve the same end with less complex testing procedures.

Lord (1968) has observed that the use of branched testing as norm-referenced measurement is not warranted under circumstances where item difficulty is not very heterogeneous. In contrast, Glaser (1967) has suggested that some form of sequential testing could prove fruitful in a program where tests are used to make instructional decisions about individuals; that is, where measurement is criterion-referenced.

THE TEST MODEL

The branched test was designed for the mathematics curriculum in IPI but is applicable to any curriculum for which an established learning hierarchy of prerequisite relationships among objectives exists. The specific unit to which the model was applied consisted of eighteen objectives in addition and subtraction typically encountered by third and fourth grade students. A hierarchy for the objectives had been hypothesized after extensive study. Validation of the hierarchy was accomplished concurrent with the study.

Figure 1 illustrates graphically the prerequisite relationships among objectives. The structure reveals that objectives 6, 17, and 18 are terminal; that is, are prerequisite to no other objectives in the unit. Two major sequences, sets of objectives whose elements are linked together in prerequisite dependencies, emerged as dominant in the structure. The sequence consisting of objectives 1, 4, 7, 10, 12, 13, 14, 16, and 17 includes strictly addition skills whereas the sequence containing objectives 2, 5, 8, 11, 15, and 18 is exclusively subtraction. Skill 6 and sequences containing it integrate the two operations of addition and subtraction.

A test model was developed which relied heavily upon the capabilities of a computer for accurate and efficient administration. The model required that

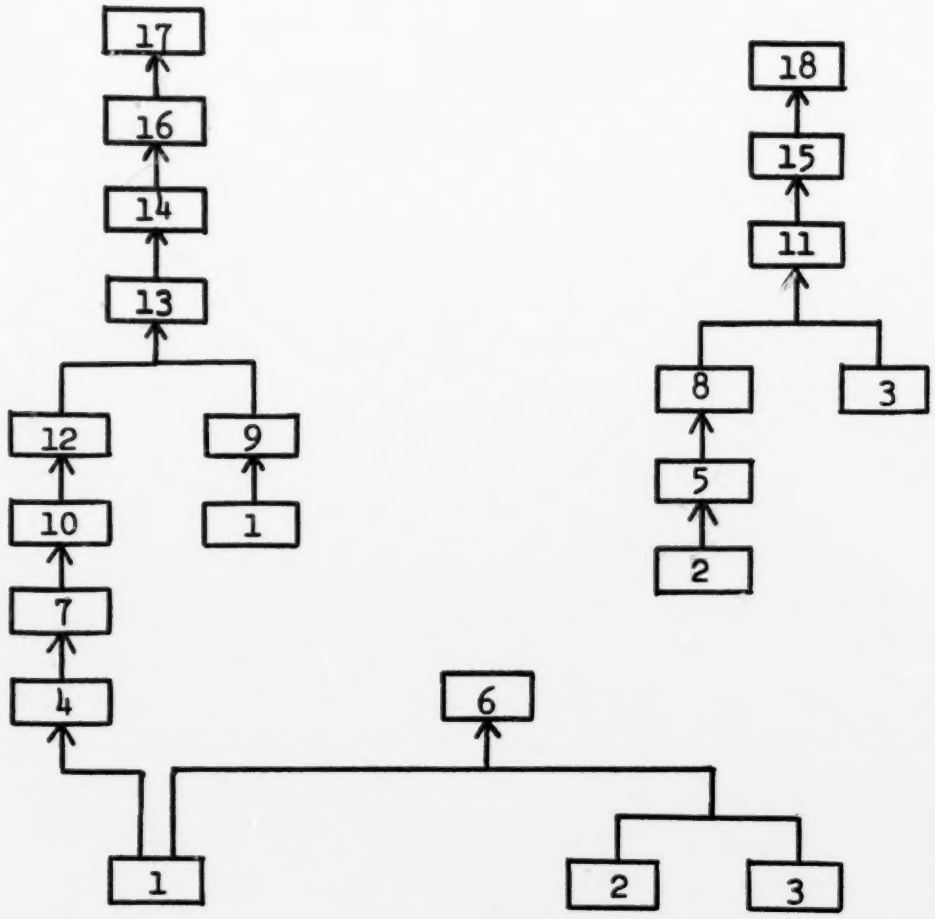

FIGURE 1

Hierarchy for Computer-Assisted Branched Testing on a Unit with 18 Objectives

the computer be programmed to control the presentation of test items for each individual. This control was exercised at two levels of decision making. At the first level, when testing a specific objective, it was necessary to determine after each response whether or not sufficient information was available from it and previous responses to decide if the examinee had proficiency in the skill. At the second level, a decision was made which governed the ordering of objectives to be tested.

Decision Making About Proficiency of Objectives.--It was both inefficient and impractical to test over the entire population of items for a specified objective since in the unit used for the study the population of items for the objectives varied in number from fifty-five to several million. Therefore, a measure of an examinee's proficiency in a specified objective was obtained by an item-sampling process which provided items for him until some decision was

reached regarding his status on the objective.

A concern for building statistical confidence into the decision process which classified an examinee as either proficient or not proficient in an objective resulted in a Bernoulli-type experiment the results of which follow a binomial distribution. The assumptions of the experiment were thus three in number. The possible number of outcomes for each trial were assumed to be two; the probability of each outcome was assumed constant over trials; and the outcome of any trial was assumed independent of the outcome of all other trials.

The model of testing assumed that at any given moment in time, a single numerical value represented the proficiency of an examinee with respect to the specified objective. His relative true score on the population of items was an estimate of this proficiency. Thus, proficiency was construed to be a parameter which was the probability of a correct response to any random item from the population.

Since the number of items required to determine an examinee's proficiency in a particular objective varied from student to student and in the interest of testing a representative sample of the population of items for an objective, items were constructed by computer as they were needed using item-generation rules. Initially, an examinee was presented with an item which was randomly generated from the population of items for the objective being tested. After the examinee responded to the item, the computer scored the response as either being correct or incorrect. At this point a decision was made which exercised one of the following options. The examinee had mastered the objective, had not mastered the objective, or had not responded to a sufficiently large sample of items to make a decision regarding mastery or non-mastery of the objective. If a decision as to mastery or non-mastery was not made, another item was generated and the process was repeated. Items were generated and scored until the process was halted according to some predetermined criteria for the maximum

number of items to be tested for a single objective.

Obviously, any sampling plan which did not exhaust the population of items may have led to an incorrect decision about the mastery of an objective. Since exhaustive testing was impossible, it was necessary to live with the risk of making wrong decisions. To define a sampling plan it was necessary to specify the maximum risks of incorrect decisions which were tolerable.

The two risks involved in making a decision regarding an individual's proficiency on an objective were the risk of requiring a prescription and work on a skill when it was not necessary (Type I) and the risk of certifying mastery of an objective when in fact a prescription and work were necessary (Type II). Errors were perceived to be of consequence in the instruction process and in the branching logic for the test which itself has implications for instruction.

A Type I error seemed to be of lesser consequence than a Type II error from the point of view of instruction and testing. The most serious error which could have resulted as a consequence of branched testing was seen to be one which led to claiming mastery for skills which were in fact not mastered. Such an error might have led to a child having difficulty proceeding through a unit and might eventually have resulted in an impasse in instruction. From the point of view of the logic of branching, an error of Type II compounded this difficulty since it resulted in indicating mastery for objectives which were not and would not be tested. A Type I error would have at worst required that the student pursue a review like study of skills in which he was already proficient.

Since it was necessary to function knowing that a Type I or Type II error could occur within the item sampling model, it was desirable to control the risks which were taken. A sampling plan which satisfied the conditions thus far specified was given by the sequential probability ratio test (Wald, 1947)

of strength (α,β) for testing the hypotheses:

1) H_o: $p = p_o$

2) H_1: $p = p_1$

In the model, 'p' was the unknown proportion of items which would have been answered incorrectly if testing had been over the entire population of items for the objective. The risks which were taken were specified in the following manner. The probability of declaring a lack of mastery for the objective should not exceed some small predetermined value α whenever $p \leq p_o$ and the probability of declaring mastery of the objective should not exceed some small value β whenever $p \geq p_1$. If 'p' was situated between p_o and p_1 no decision was made and thus no error occurred. It becomes clear that control of the decision process was a function of four numbers, p_o, p_1, α, and β, all of which were parameters which could be varied by the test constructor. The choice of these values was based on considerations relative to the testing and instructional phases of IPI and varied for different objectives.

Values for p_o and p_1 for this study were selected after consultation with curriculum experts who were familiar with the unit. Alpha and beta were set at .20 and .10, respectively. p_o and p_1 were set according to the particular objective. For the majority of the objectives, p_o was set at .15 and p_1 at .40.

In actual operation, one could say that the probability of declaring non-mastery for most of the objectives did not exceed .20 whenever $p \leq .15$ and the probability of declaring mastery of the objective did not exceed .10 whenever $p \geq .40$. The test for deciding mastery or non-mastery of an objective was described as follows:

Let x_i represent the evaluation of the response to the $i^{\underline{th}}$ item where $x_i \in U$ and

$$U = \begin{cases} 0 \text{ if the } i^{\underline{th}} \text{ item was answered correctly} \\ 1 \text{ if the } i^{\underline{th}} \text{ item was answered incorrectly} \end{cases}$$

Since 'p' was the proportion of items answered incorrectly in the population of items, the probability of getting a sample equal to $(x_1, x_2, \ldots, x_m)$ was $p^{w_m} (1 - p)^{r_m}$ where w_m was the number of items in the sample of size 'm' answered incorrectly and $r_m = m - w_m$.

Under H_o: $p = p_o$, the probability became $p_o^{w_m} (1 - p_o)^{r_m}$ and under H_1: $p = p_1$, the probability became $p_1^{w_m} (1 - p_1)^{r_m}$. The sequential probability ratio test was then applied and an acceptance number (a_m) and rejection number (u_m) which were dependent upon the values of α, β, p_o, p_1, and m were computed. Testing continued if $a_m < w_m < u_m$. If $w_m \geq u_m$, non-mastery was indicated and if $w_m \leq a_m$ mastery was indicated.

Decision Making Related to Branching.--Once a decision was reached about an individual's proficiency on a particular objective, he was branched for testing on another objective. Inspection of the unit hierarchy revealed the existence of the seven sequences of objectives found in Table 1. Each sequence was comprised of a set of objectives which were ordered such that starting from the left, each objective was the prerequisite of all objectives to its right.

TABLE 1

List of Sequences Based on the Hierarchy
for the 18 Objective Unit

Sequence	Objectives Comprising the Sequence
1	1, 4, 7, 10, 12, 13, 14, 16, 17
2	1, 9, 13, 14, 16, 17
3	1, 6
4	2, 6
5	3, 6
6	2, 5, 8, 11, 15, 18
7	3, 11, 15, 18

A plan for branching was devised using the rationale than an examinee who had evidenced proficiency in an objective with almost no incorrect solutions should be branched to a more difficult objective than an examinee who had mastered the objective with several incorrect solutions. Likewise, it was believed that the examinee who did not have mastery of an objective and who answered nearly all of the items incorrectly should be branched to an easier objective than the examinee who had no mastery but responded correctly to a larger proportion of the items presented. Thus, the following branching plan was devised.

During the testing process, $PR_{i,j}$, the percentage of items answered correctly by individual 'i' on objective 'j' was determined by the computer. If the objective was mastered and $PR_{i,j} \geq (1 - .5p_o)$ he was branched to the most difficult untested objective in the sequence whereas if $PR_{i,j} < (1 - .5p_o)$ he was branched to an objective not yet tested in a move best described as a binary branch. In short, he was branched to a more difficult objective midway between those not already tested.

In the event that the examinee failed to master the objective, a similar procedure was used to branch him to a less difficult objective in the sequence. In this case, if $PR_{i,j} < .5(1 - p_o)$ he was branched to the least difficult untested objective of the sequence. Whenever $PR_{i,j} \geq .5(1 - p_o)$ he was branched to a less difficult objective midway between those not already tested.

Testing for all examinees began with objective twelve of sequence one. The branch and test cycle continued until a judgment was made about each objective of the sequence, whereupon another branch resulted in testing of a sequence for which decisions about each objective had not been made.

PROCEDURES

The branched test was administered to a sample of seventy-five students in grades one through six at the Oakleaf Elementary School. On two separate

occasions each student was given the computer test using a teletypewriter. In most cases the tests were taken by each student on consecutive days. In no instance did an examinee have instruction on the unit between tests. Since items were constructed using a random number generator, each test was unique but parallel to all others.

Since there was likely to be a marked variation in the branching routes and test characteristics for individuals at the extremes of the proficiency continuum, two groups, each with ten students, were included in the test sample. They will be referred to as the low and high proficiency groups. An additional fifty-five students formed the middle proficiency group. The latter group was comprised of children with a wide range of proficiencies. Students were assigned a priori to the groups by the IPI coordinator at Oakleaf School. An effort was made to include students of varying experience with the unit within the middle proficiency group. Of the seventy-five students tested, twenty-eight had not yet entered the unit, eleven were working in the unit, and thirty-six had completed the unit at an earlier date.

Since the tests were presented by teletypewriter, a complete record of each test was preserved. As the examinee worked at the teletypewriter, a record of each objective tested was maintained. After completing the branched test he was required to take a paper and pencil test on all of those objectives not included on the branched test. Thus, a measure of his proficiency in every objective in the unit was recorded.

RESULTS AND DISCUSSION

Individual test profiles resulting from the study were used to ascertain the validity of the unit's structure. Examination of these profiles revealed a minimum number of inconsistencies in the profiles. An inconsistency was defined as an objective being unmastered while an objective to which it was prerequisite was mastered. Of the possible number of inconsistencies, ninety-nine

percent did not occur. With a valid structure affirmed, meaningful consideration could be given to the test results.

The branched test was highly endowed with content validity as the skills were behaviorally defined and thus precisely translated into item-generation rules. From another point of view, the test could be considered valid only if it reflected an accurate measure of the examinee's proficiencies. Since inferences were made about objectives which were not tested, a second validity concern was with the accuracy with which the branched test predicted an examinee's performance on objectives for which testing did not occur.

To obtain a measure of the extent to which the branched test had predictive validity in the sense described above, an index was defined which revealed the average proportion of objectives correctly inferred as mastered or unmastered by the branched test. Such an index was possible since all objectives not tested at the computer were tested by paper and pencil. An index of predictive validity was determined by calculating the proportion of correct inferences for each examinee's profile and then computing the mean of the proportions for the entire sample. Such an index was determined for both the first and second administrations of the test. The index for both testings was .99.

Assuming that the test and structure were valid, the question of reliability for the branched test was approached by examining the extent to which identical placement profiles for the unit were obtained from two administrations of the test. A necessary assumption was that no instruction involving the unit occurred between tests. One approach to the problem was to assign a score to each of the seven sequences on the basis of the computer test profiles. If, for example, the examinee had mastery of objectives 1, 4, 7, and 9 of sequence one and no mastery for the remaining objectives in the sequence, he was assigned a score of four for sequence one. Once this was repeated for all seven sequences on each of the tests, the scores on the first test of each sequence were correlated with

the scores of the corresponding sequence obtained from the second testing for each individual. For both the low and high proficiency groups, the correlation coefficients were 1.00 for all seven sequences. In each case, the coefficients were determined with an N of ten. The coefficients for the middle proficiency group, N = 55, are reported in Table 2.

TABLE 2

Correlation Coefficients Between Repeated Measures of the Seven Sequence Hierarchy for the Middle Proficiency Group

N	Sequence	r_{12}
	1	.95
	2	.90
	3	.83
55	4	.81
	5	.91
	6	.96
	7	.96

To obtain a relative measure of the consistency of the entire test from one testing to another, a reliability index was defined and determined in much the same manner as the validity indices described in the preceding discussion. The profiles which were compared for inconsistencies in this case were those obtained from the computer test. The number of errors observed by comparing the first computer test profile with the second were counted and the proportion of the number of inconsistencies found to the number which were possible was determined. This done for all seventy-five examinees, a reliability index was defined as the mean of the proportions. The index for the entire sample was .96. A reliability index of .96 reflects that of the inconsistencies which could have occurred from test I to test II for all examinees, 96% did not occur.

Some of the most interesting results of the study became apparent when the branched test was compared with the conventional paper and pencil test

currently used in the instructional program. The mean time required to complete a conventional test on the objectives in the unit was approximately seventy-five minutes. Table 3 provides the mean rate in minutes for the three proficiency groups for the first and second test administrations. The table reveals that the middle group, which is most nearly representative of the group who take the conventional test, required less than one-half as much time to complete the branched test as to complete the conventional test.

TABLE 3

Mean Rate in Minutes for Groups of Varying Proficiency to Complete the Branched Test

Group Proficiency	Test	N	Mean Rate	Variance
Low	I	10	12.50	7.08
	II		10.70	7.78
Middle	I	55	36.07	175.30
	II		34.62	197.68
High	I	10	17.40	14.82
	II		13.20	6.76
Pooled	I	75	30.44	220.23
	II		28.57	247.75

To account for the extreme differences in rates one need only observe that the mean number of objectives tested on the branched test was substantially less than on the conventional test. On the conventional test every examinee was presented a fixed number of items on all of the eighteen objectives. For the three proficiency groups pooled, the mean number of objectives tested using the branching model were 7.11 and 6.99 on test I and test II, respectively.

In the case of the low proficiency group, every child was tested on precisely seven objectives, the minimum number of objectives possible for an examinee having mastery of none of the eighteen objectives in the unit. For

this group, the routing sequence which every examinee followed on both tests was 12-4-1-10-3-5-2.

For the high proficiency group, each of the ten examinees had profiles which indicated mastery for all eighteen objectives on both tests. Each examinee required testing on five objectives. This was the minimum number of objectives on which testing was possible. The routing for these individuals was the set 12-17-11-18-6.

For the group with middle proficiency, a mean of 7.4 objectives were tested. Since this group was most like the target population for which the test was designed, it is anticipated that branched testing could eliminate testing on an average of about 10.6 of the eighteen objectives.

The branching design within the test was constructed so that the lower and upper bound for the number of objectives which could be tested were five and ten, respectively. Thus, the examinee whose profile was most difficult to complete was tested on ten objectives or only 55% of the objectives on which he would be tested in the traditional testing program. The individual who had mastery of the unit with testing on just five objectives was tested on only 28% of the objectives required by the conventional tests. The range of the number of objectives tested for examinees in the study was from five to ten.

A function of the item-sampling procedure and the reduction in the number of objectives tested, the mean number of items which required testing on the branched test was substantially less than the 150 items required on the fixed length conventional test. For all groups combined, an average of 52.12 items was required per branched test. As noted before, fewer items were required to declare non-mastery than to declare mastery of an objective. An examinee in the low proficiency group who was unable to respond correctly to a single item on his branched test would have been presented fourteen items since it required two consecutive incorrect responses per objective to certify non-mastery. The

fewest number of items which could be presented to an examinee who had proficiency of all eighteen objectives was thirty-three.

The routing plan used in this study was but one of many which could have been implemented. Two other branching techniques were simulated using the placement profiles which resulted from the computer-assisted branched test. The purpose of the simulation was to determine if either routing method would reduce the number of objectives to be tested while still arriving at the same unit profile. The branching rule for the first technique was to branch up one objective in the sequence if the objective currently being tested was mastered and down two objectives in the sequence if it was not. The second technique reversed the magnitude of the steps taken with the first technique.

The results of the simulation show that the method used for routing in the study was markedly superior to either of the methods to which it was compared. In 150 trials, the first method required testing for fewer objectives only eleven times, the same number of objectives forty-seven times, and more objectives in ninety-two cases. The second method was better than the first simulated procedure but was still not as good as the one used in the study. It required fewer objectives thirty-six times, the same number of objectives twenty-four times, and more objectives ninety times.

The routing approach used in the study was to test every examinee on objectives which were found in the middle of the major sequences of the structure. Thus, every child was tested on objectives eleven and twelve. Branching after the initial objectives in each sequence had been tested depended upon the level of proficiency at which a decision on the objective currently being tested had been made. In the course of testing, fifty different branching routes were followed. The latter clearly established the flexibility of this branching process in adapting to individual differences.

The branched test implemented in this study performed its function well. Perhaps the most dramatic of all conclusions reached was with regard to the impact an extensive testing program such as this could have on instruction in IPI mathematics. During the course of a school year, large numbers of hours now spent in testing could be invested in instructional activities or in supplementary diagnostic testing.

The typical expression of reservation regarding branched testing has been with regard to its characteristic inability to improve upon conventional test measurement for the examinee of average ability. The results of this study strongly suggest that the branched test was extremely effective in tailoring measurement to the group with middle proficiency. To a large extent, this can be attributed to the specific unit of work which was tested. For units with fewer objectives and with smaller sequences, the affect of branched testing for the majority of examinees may be less pronounced.

Nevertheless, this study has shown that criterion-referenced branched testing can be used effectively for all students in an individualized instruction setting. Further, the measures yielded by such a procedure can be at least as valid and reliable as those for a conventional test with the additional bonus that they are obtained in less time with testing of fewer items and objectives.

REFERENCES

Angloff, W. H. and Huddleston, E. M. "The Multi-Level Experiment. A Study of a Two-Level Test System for the College Board Scholastic Aptitude Test," Educational Testing Service Statistical Report, 58-21. Princeton, New Jersey: Educational Testing Service, 1958.

Bayroff, A. G. and Seeley, Leonard C. "An Exploratory Study of Branching Tests," United States Army Behavioral Science Research Laboratory Technical Research Note, 188. Washington, D.C.: Military Research Division, 1967.

Cleary, T. Anne; Linn, Robert; and Rock, Donald. "An Exploratory Study of Programmed Tests," Education and Psychological Measurement. XXVIII, 1968, pp. 347-349.

Glaser, Robert. "Adapting the Elementary School Curriculum to Individual Performance," Proceedings of the Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1967.

Hanson, Duncan N. and Schwarz, Guenter. "An Investigation of Computer-Based Science Testing," Report submitted to the College Entrance Examination Board. Tallahassee, Florida: The Florida State University, 1968.

Lord, Frederick. "Some Test Theory for Tailored Testing," Report to the conference on Computer-Based Instruction, Learning, Testing, and Guidance. Austin, Texas, 1968.

Wald, Abraham. Sequential Analysis. New York: John Wiley and Sons, 1947.

Waters, Carrie Jean. "Preliminary Evaluation of Simulated Branching Tests," United States Army Personnel Research Office Technical Research Note, 140. Washington, D.C., 1964.